2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669383

Wasserstein-Distance-Based Multi-Source Adversarial Domain Adaptation for Emotion Recognition and Vigilance Estimation

Yun Luo
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
angeleader@sjtu.edu.cn

Bao-Liang Lu*
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
bllu@sjtu.edu.cn

***Abstract*—To build a subject-independent affective model based on electroencephalography (EEG) is a challenging task due to the domain shift problem caused by individual differences in EEG data. In this paper, we prove a new generalization bound based on Wasserstein distance for multi-source classification and regression problems. Based on our bound, we propose two novel Wasserstein-distance-based multi-source adversarial domain adaptation methods (wMADA) for learning domain invariant and task discriminative domain mappings by dynamically aligning different domain mappings. We evaluate our methods on two typical EEG datasets. The experimental results demonstrate that our wMADA methods successfully handle the multi-source domain shift problem in creating subject-independent affective models and outperform the state-of-the-art domain adaptation methods.**

***Index Terms*—Affective brain-computer interface, EEG-based emotion recognition, EEG-based vigilance estimation, multi-source domain adaptation**

I. Introduction

A major obstacle for applying aBCIs to the real-world scenarios is the structural variability of electroencephalography (EEG) signals between different subjects, which causes the domain shift problem and can not make a model trained by a subject generalize well to another subject.

Domain adaptation (DA) is one of the promising ways to dealing with the domain shift problem [1]. DA assumes that the marginal distribution of the labeled source domain is different from the unlabeled target domain while their conditional distributions are the same. DA methods alleviate the domain shift problem by mapping the two domains into a common feature space where the marginal distributions of these two domain mappings are similar. In recent years, researchers have successfully built subject-independent EEG-based emotion recognition models for aBCIs by applying single-source DA methods [2]–[5]. And most of them consider all the source subjects as one domain. However, different subjects usually have different marginal distributions. As a result, typical single-source DA methods may lead to suboptimal results when tackling these multi-source tasks.

In this paper, we view data from different subjects belong to different domains and develop a multi-source subject-independent approach to overcoming the domain shift problem in two common aBCIs paradigms: EEG-based emotion recognition and vigilance estimation. We give a new generalization bound for both classification and regression problems which have multiple source domains and propose a novel DA method called Wasserstein-distance-based multi-source adversarial domain adaption (wMADA). Here, we introduce two versions of wMADA: wMADA-α and wMADA-β. wMADA-α method directly minimizes the bound without considering the relationship between different source domain mappings. wMADA-β method minimizes the bound while aligning multi-source domain mappings by introducing a public discriminator.

*Corresponding author
978-1-6654-0126-5/21/$31.00 ©2021 IEEE

II. Methodology

A. Theoretical Analysis

Here we use the dual form of Wasserstein-1 distance [6]:

$$W_1(\mu_S, \mu_T) = \frac{1}{K} \sup_{||f||_L \leq K} \mathbb{E}_{x\sim\mu_S}[f(x)] - \mathbb{E}_{x\sim\mu_T}[f(x)], \quad (1)$$

where $||f||_L \leq K$ denotes the set of all *K*-Lipschitz continuous functions $f : \Omega \to \mathbb{R}$.

Theorem 1: [7] Let $\mu_S,\ \mu_T \in \mathcal{P}(\Omega)$ be two probability measures. Assume $\mathcal{H}$ is the hypothesis class of K-Lipschitz continuous functions with some certain K, then we have:

$$\epsilon_T(h) \leq \epsilon_S(h) + 2KW_1(\mu_S, \mu_T) + \lambda, \quad (2)$$

where λ is the combined error of the optimal hypothesis h^* minimizes $\epsilon_S(h) + \epsilon_T(h)$.

Shen *et al.* [7] proposed a learning bound for DA with Wasserstein distance in the single source case. Here we broaden the bound of Eq. (2) for the multi-source cases.

Theorem 2: Let $\mathcal{H}$ be a hypothesis class with K-Lipschitz continuous functions , and $\{S_i\}_{i=1}^k$ and T are k source

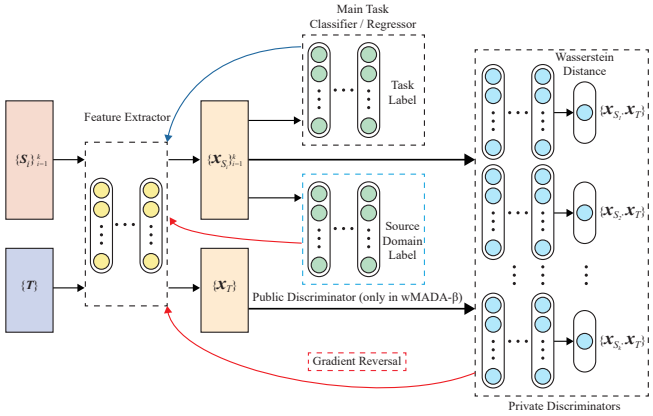


Fig. 1. The architecture of the proposed wMADA-α and wMADA-β methods. The two methods have the similar network architecture except for that wMADA-α does not have a public discriminator. For the wMADA-α method, all the components are trained simultaneously. For the wMADA-β method, in the pre-training step, all the components except the private discriminators are trained; in the self-adaptive training step, all the components are trained together. Specially, we apply adversarial training with gradient reversal.

domains and the target domain, then, $\forall \alpha_i \geq 0,\ \sum_{i=1}^{k} \alpha_i = 1$, $\forall h \in \mathcal{H}$, we have:

$$\epsilon_T(h) \leq \sum_{i=1}^{k} \alpha_i(\epsilon_{S_i}(h) + 2KW_1(\mu_{S_i}, \mu_T)) + \lambda_\alpha, \quad (3)$$

where λ_α is the error of the optimal hypothesis on the mixture source domain $\sum_{i=1}^{k} \alpha_i S_i$.

Proof 2.1: Denote $\tilde{S} = \sum_{i=1}^{k} \alpha_i S_i$ as the mixed source domain weighted by $\{\alpha_i\}_{i=1}^{k}$ of the k source domains, whose distribution is $\mu_{\tilde{S}} = \sum_{i=1}^{k} \alpha_i \mu_{S_i}$. From **Theorem 1**, we have:

$$\epsilon_T(h) \leq \epsilon_{\tilde{S}}(h) + 2KW_1(\mu_{\tilde{S}}, \mu_T) + \lambda_\alpha, \quad (4)$$

and we can upper the bound of $2KW_1(\mu_{\tilde{S}}, \mu_T)$ as follows,

$$\begin{aligned} 2KW_1(\mu_{\tilde{S}}, \mu_T) &= \sup_{||f||_L \leq 2K} E_{x\sim\mu_{\tilde{S}}}[f(x)] - E_{x\sim\mu_T}[f(x)] \\ &\leq \sum_{i=1}^{k} \alpha_i \sup_{||f||_L \leq 2K} E_{x\sim\mu_{S_i}}[f(x)] - E_{x\sim\mu_T}[f(x)] \\ &= \sum_{i=1}^{k} \alpha_i 2KW_1(\mu_{S_i}, \mu_T), \end{aligned} \quad (5)$$

Replacing $\epsilon_{\tilde{S}}(h)$ with $\sum_{i=1}^{k} \alpha_i \epsilon_{S_i}(h)$, we complete the proof.

Redko *et al.* discussed the convergence of empirical Wasserstein distance to its real distance [8]. The bound here is suitable for both classification and regression problems.

B. wMADA-α

1) Framework: Inspired by our theoretical results, we first propose wMADA-α, which is illustrated in Figure 1. Suppose we sample from k labeled source domains $\{S_i\}_{i=1}^{k}$ and one unlabeled target domain $\{T\}$. The last term in the generalization bound of Eq. (3) can be ignored under the assumption of DA. Namely, we only need to minimize the training error of source domains and the Wasserstein distance between each source and target domain.

Our proposed wMADA-α method consists of a common feature extractor, a main task classifier or regressor, and k private discriminators. The feature extractor aims to map all the domains to a common space. We denote the i-th source mapping as x_{S_i} and target mapping as x_T. The private discriminators, which align the source domain mapping with target domain mapping, calculate the empirical Wasserstein distance between the marginal distribution of each source domain mapping and target domain mapping in an adversarial manner. We use gradient reversal [9] to realize adversarial training.

2) Loss: In the training phase, all the source domains and target domain are first mapped into a common space with the common feature extractor. We denote our main task loss as $\mathcal{L}_{main_{S_i}}$ for each source domain based on our main task, where S_i denotes source domain i. If the main task is a classification problem, $\mathcal{L}_{main_{S_i}}$ is the sum of typical cross-entropy loss of each source domain. If it is a regression problem, $\mathcal{L}_{main_{S_i}}$ is the sum of root mean squared error. We denote θ_f and θ_F as the parameters of the main task network and common feature extractor, respectively.

Next, we need to calculate the Wasserstein distance between different domain mappings. The output of each private discriminator is denoted as $d_i(x)$ which can be used to calculate the empirical Wasserstein distance between x_{S_i} and x_T. With the private discriminators, we aim to align each x_{S_i} with x_T. In this way, we can minimize the empirical Wasserstein distance between x_T and x_{S_i} in an adversarial way [6],

$$\min_{\theta_F} \max_{\theta_{d_i}} \quad \mathbb{E}_{x\sim x_T}[d_i(x)] - \mathbb{E}_{x\sim x_{S_i}}[d_i(x)], \quad (6)$$

where θ_{d_i} represents the parameters of $d_i(x)$, and we define the distance as:

$$\mathcal{L}_{w_i} = \mathbb{E}_{x\sim x_T}[d_i(x)] - \mathbb{E}_{x\sim x_{S_i}}[d_i(x)]. \quad (7)$$

For the constraint of K-Lipschitz continuousness, we use gradient penalty [10] work to make the training process more stable. The gradient penalty is as follows,

$$\mathcal{L}_{grad_i} = \lambda(||\nabla_{\hat{x}} d_i(\hat{x})||_2 - 1)^2, \quad (8)$$

where $\hat{x}$ is the random linear interpolation between x_{S_i} and x_T, namely $\hat{x} = \beta x_T + (1-\beta)x_{S_i}$ for some random β sampled between 0 and 1.

According to the proposed generalization bound, the loss on each domain will have a weight α_i from the factors of a convex combination. Here we define the weight as:

$$\alpha_i = \frac{\exp(\mathcal{L}_{main_{S_i}} + \mathcal{L}_{w_i})}{\sum_{j=1}^{k} \exp(\mathcal{L}_{main_{S_j}} + \mathcal{L}_{w_j})}. \quad (9)$$

We use this weight for two reasons: (a) this weight could lead to a brief upper bound [11]; (b) it adaptively corresponds to the main task loss and Wasserstein distance between different domain mappings, and we give the following weight rule:

the larger the loss and distance, the heavier the weight. The larger loss means that the corresponding main task network should be fully trained, so we need to give it a larger weight. Besides, larger distance means a larger domain shift between target and source domain mapping, and we need to use a larger weight to reduce the distance.

The updating rule during the training is:

$$\min_{\theta_f,\theta_F} \sum_{i=1}^{k} \alpha_i (\mathcal{L}_{main_{S_i}} + \max_{\theta_{d_i}} \mathcal{L}_{w_i}) + \sum_{i=1}^{k} \min_{\theta_{d_i}} \mathcal{L}_{grad_i}. \quad (10)$$

C. wMADA-β

1) Framework: We now propose wMADA-β method which aligns the target domain mapping with the multi-source domain mappings, and aligns source domain mappings with each other simultaneously. The wMADA-β method illustrated in Figure 1 has an extra public discriminator, which aligns multi-source domain mappings with each other by constraining all of the source domain mappings to one distribution with adversarial training. The functions of the other components are the same as the wMADA-α method. To better align different domain mappings, the wMADA-β method has two training steps: pre-training and self-adaptive training.

2) Loss: In the pre-training phase, all the source domains and target domain are first mapped into a common space with the common feature extractor. The loss $\mathcal{L}_D$ for the public discriminator is the cross-entropy for domain classification. And other notations are the same as the wMADA-α method. During this phase, we update the main task network and the public discriminator according to the following rule:

$$\min_{\theta_f,\theta_F} \sum_{i=1}^{k} \mathcal{L}_{main_{S_i}} + \min_{\theta_D} \max_{\theta_F} \mathcal{L}_D, \quad (11)$$

where θ_D represents the parameters of the public discriminator. By training $\mathcal{L}_D$ in an adversarial manner, we hope the common feature extractor can fool the public discriminator and make it more powerful. This process can finally reach Nash Equilibrium, in which all source domain mappings will have a similar marginal distribution. Namely, the multi-source domain mappings are aligned in the first place.

In the self-adaptive training phase, we need to minimize the bound in two aspects: (a) minimize the main task loss and Wasserstein distance between target and different source domain mappings by an adaptive weight; (b) train the public discriminator in the same way described in the pre-training phase to align the multi-source domain mappings. To better align different domain mappings, not only do we need to train the public discriminator to keep the similarity of different source domain mappings, but also we need to use an adaptive weight which both considers the relationship between source and target domain mappings and the relationship between multi-source domain mappings.

The output of the public discriminator is denoted as $D(x)$ which is a *k*-dimensional vector. It indicates the probability that the input x belongs to the *k* source domain mappings. We define the relative distance from x_{S_i} to the other source domain mappings as follows:

$$s_i = \frac{\exp(KL(D(x_{S_i})||\boldsymbol{U}))}{\sum_{j=1}^{k} \exp(KL(D(x_{S_j})||\boldsymbol{U}))}, \quad (12)$$

where $\boldsymbol{U}$ is the uniform distributed matrix whose shape is like $D(x_{S_i})$ and elements are all $1/k$. We use a softmax operation on the *KL*-divergence between the public discriminator's output and a uniform distributed matrix which can indicate the relative position among the source domain mappings. When a source domain mapping is 'further' from other source domain mappings, its s_i will be larger.

And now we can define the adaptive weight as:

$$\alpha_i = \frac{\exp(\mathcal{L}_{main_{S_i}} + \mathcal{L}_{w_i} + s_i)}{\sum_{j=1}^{k} \exp(\mathcal{L}_{main_{S_j}} + \mathcal{L}_{w_j} + s_j)}. \quad (13)$$

Our intuition is, if x_{S_i} has a higher main task error, a higher distance from x_T and is further from the other source domain mappings, we give it a higher weight. Comparing with the wMADA-α method, we introduce an s_i factor, which ensures to align multi-source domain mappings with each other while aligning target domain mapping with source domain mappings. Besides, the weight is according with the weight rule mentioned in the wMADA-α method, the larger the loss and the distance, the larger the weight.

So the updating rule during the self-adaptive training phase is:

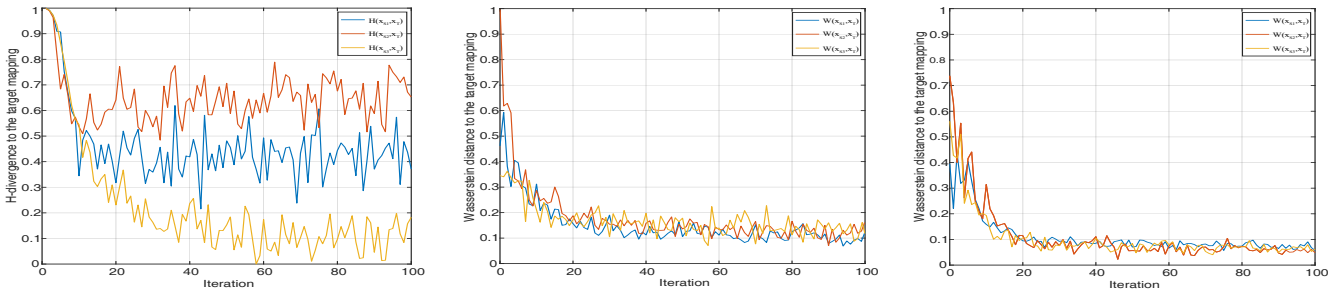
$$\begin{aligned} &\min_{\theta_f,\theta_F} \sum_{i=1}^{k} \alpha_i (\mathcal{L}_{main_{S_i}} + \max_{\theta_{d_i}} \mathcal{L}_{w_i}) + \sum_{i=1}^{k} \min_{\theta_{d_i}} \mathcal{L}_{grad_i} \\ &+ \min_{\theta_D} \max_{\theta_F} \mathcal{L}_D. \end{aligned} \quad (14)$$

III. Experiments

A. Datasets and Data Pre-processing

In this work, we evaluate different DA and our proposed methods on two typical EEG datasets, SEED[1] [12] and SEED-VIG[2] [13]. The SEED dataset contains the EEG signals of 15 participants. They were required to watch 15 well-prepared video clips that can elicit exactly one of the three kinds of emotion: positive, neutral, and negative. The SEED-VIG dataset consists of the EEG signals and electrooculography (EOG) signals recorded from 23 subjects. They were required to drive in a simulated driving system for two hours to elicit different vigilance state. For the SEED dataset, differential entropy (DE) feature [14] has been extracted from 5 frequency bands. For the SEED-VIG dataset, we use forehead EOG and EEG signals. We extract the same feature following the existing studies [4].

[1]http://bcmi.sjtu.edu.cn/∼seed/index.html
[2]http://bcmi.sjtu.edu.cn/ seed/download.html



(a) (b) (c)

Fig. 2. Comparison of $\mathcal{H}$-divergence and Wasserstein distance. (a) Change of $\mathcal{H}$-divergence in MDAN. (b) and (c) Changes of Wasserstein distance in wMADA-α and wMADA-β, respectively.

Methods	SEED		SEED-VIG		SEED-VIG	
	Acc.	Std.	PCC	Std.	RMSE	Std.
SVM/SVR	0.582	0.139	0.761	0.231	0.169	0.067
TCA	0.640	0.147	0.779	0.215	0.160	0.054
DANN	0.792	0.131	0.840	0.154	0.143	0.059
ADDA	0.812	0.061	0.844	0.134	0.141	0.051
WGANDA	0.866	0.047	0.852	0.098	0.154	0.053
M^3SDA	0.868	0.053	0.852	0.081	0.141	0.055
MDAN	0.868	0.042	0.859	0.112	0.140	0.054
wMADA-α	0.880	0.045	0.869	0.104	0.139	0.050
wMADA-β	**0.893**	**0.040**	**0.891**	**0.069**	**0.139**	**0.048**

TABLE I
PERFORMANCE OF DIFFERENT DA METHODS.

B. Evaluation Details

We use leave-one-subject-out cross-validation, which is a widely applied evaluation criterion in the existing subject-independent aBCIs [2]–[5], to demonstrate the effectiveness of the proposed methods. Specifically, we leave one subject as the target domain for each time, and other subjects (14 subjects for SEED, 22 subjects for SEED-VIG) are regarded as source domains.

The support vector machine (SVM)/support vector regressor (SVR) is selected as the baseline method. We also compare the results of TCA [15], DANN [9], ADDA [16], and WGANDA [5]. For multi-source DA methods, we introduce two representative methods to aBCIs: M^3SDA [17] and MDAN [11].

IV. RESULTS

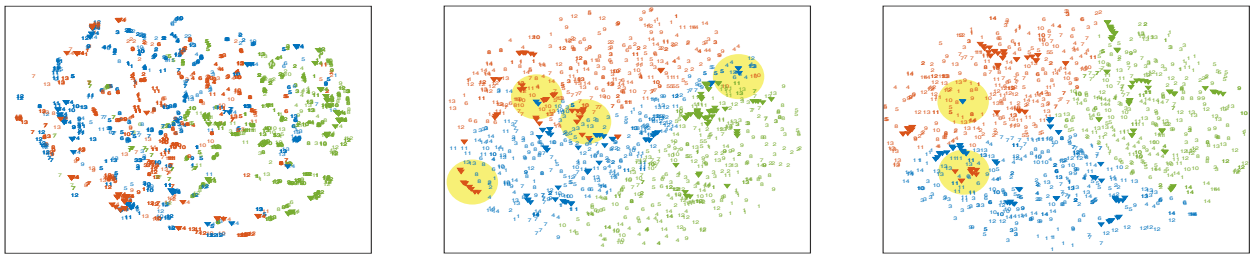
Figure 2 shows the normalized distances between the three randomly selected source domain mappings and the target mapping on the SEED dataset during the training of MDAN [11], wMADA-α and wMADA-β methods. We pre-train the private discriminators to make their outputs at the first iteration could represent the Wasserstein distance. We can see that Wasserstein distance is superior to $\mathcal{H}$-divergence, and converges much more stable. Besides, we can find that our methods get a relatively smaller distance, which implies the two corresponding domain mappings from our methods have a more similar marginal distribution. This observation also agrees with the previous theories [8] and results [6], [7].

Compared with wMADA-α, the distances in wMADA-β are closer to each other, which implies that the distances between target domain and different source domain mappings are similar. Different source domain mappings have a similar distribution and the public discriminator successfully aligns multi-source domain mappings with each other while the target and source domain mappings are aligned. Besides, the converge values of the distances in wMADA-β are also smaller than wMADA-α, which implies that the generalization bound could approximate to its minimum when the source domain mappings are aligned. Moreover, Figure 2(c) also shows that the public discriminator can improve stability during the training process.

Table I shows the performance of different DA methods on the two datasets. For the SEED dataset, we use the mean classification accuracy of all folds to estimate the performance of different methods. We list the results of SVM, TCA, and DANN from the existing work [3]. We implement ADDA, WGANDA, M^3SDA, and MDAN on the SEED dataset. Comparing with SVM, we see that DA methods significantly improve the performance. Besides, multi-source DA methods have higher accuracies. And our proposed wMADA-β method outperforms the existing methods with a mean accuracy of 89.3%.

Since the SEED-VIG dataset is a dataset for the regression problem, we use two measures, Pearson's correlation coefficient (PCC) and root-mean-square error (RMSE) to estimate the performance of different DA methods. The results of SVR, TCA, DANN, and ADDA are from the existing work [4]. And we implement WGANDA, M^3SDA, and MDAN on the SEED-VIG dataset. For PCC, multi-source DA methods have better performance. The proposed wMADA-β method achieves the best mean PCC of 0.891 among nine approaches. For RMSE, our wMADA-β method also reaches the best performance with a mean RMSE of 0.139. Moreover, wMADA-β is superior to wMADA-α on the two datasets because the multi-source domain mappings are aligned in this method.

As illustrated in Figure 3, the domain mappings produced by different methods are visualized in a two-dimensional way by using t-SNE to explain the effectiveness of the proposed methods. From Figure 3(a), we see that the domain mappings



(a) (b) (c)

Fig. 3. Two-dimensional visualization of source and target mappings from different DA methods on the SEED dataset: (a) TCA; (b) MDAN; (c) wMADA-β. Here, different colors represent different emotions, different numbers represent different source domain mappings, and triangles represent the data from target domain mapping. Note that the yellow circles shown in (b) and (c) denote overlapping areas between different emotions, and the overlapping areas in (a) are omitted because there are too many overlapping areas.

are mixed up, which indicates the traditional single-source DA methods can not work efficiently for the multi-source tasks. Figure 3(b) illustrates the domain mappings generated by MDAN. Although the mappings of different domains are clustered, there exist four areas with overlapping between different emotions. As we can see from Figure 3(c), not only does different emotion has more clear classification boundary, but also different target and source domain mappings are distributed uniformly in the space in comparison with Figure 3(b). This observation is also consistent with Figure 2 since wMADA-β method aligns the target domain mapping with the source domain mappings and aligns the multi-source domain mappings simultaneously. wMADA-β method successfully finds a space where domain mappings are domain invariant and task discriminative. And different domain mapping has a similar marginal distribution. Therefore, the multi-source domain shift problem has been handled.

V. Conclusions

In this paper, we have proven a new generalization bound based on Wasserstein distance for multi-source DA on both classification and regression problems. Based on the new bound, we have proposed the wMADA-α method for dealing with the multi-source domain shift problem and building subject-independent aBCIs models. And we also have proposed wMADA-β by aligning the multi-source domain mappings while aligning the target domain mapping with the source domain mappings. We have evaluated the performance of our methods and compared our methods with other state-of-the-art single-source and multi-source DA methods by conducting leave-one-subject-out cross-validation on two public EEG datasets. The wMADA-β method has the best performance with a mean accuracy of 89.3% on the SEED dataset and with a mean PCC of 0.891 and a mean RMSE of 0.139 on the SEED-VIG dataset. And the experimental results are in consistent with our theories.

Acknowledgments

This work was supported in part by grants from the National Natural Science Foundation of China (Grant No. 61976135), SJTU Trans-Med Awards Research (WF540162605), the Fundamental Research Funds for the Central Universities, and the 111 Project.

References

[1] M. Long, Y. Cao, Z. Cao, J. n. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. PAMI*, vol. 41, no. 12, pp. 3071–3085, Dec 2019.

[2] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *IJCAI'16*, 2016, pp. 2732–2738.

[3] H. Li, Y.-M. Jin, W.-L. Zheng, and B.-L. Lu, "Cross-subject emotion recognition using deep adaptation networks," in *Neural Information Processing*, 2018, pp. 403–413.

[4] H. Li, W.-L. Zheng, and B.-L. Lu, "Multimodal vigilance estimation with adversarial domain adaptation networks," in *IJCNN*, 07 2018, pp. 1–6.

[5] Y. Luo, S.-Y. Zhang, W.-L. Zheng, and B.-L. Lu, "WGAN domain adaptation for EEG-based emotion recognition," in *Neural Information Processing*, 2018, pp. 275–286.

[6] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[7] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *AAAI'18*, 2018, pp. 4058–4065.

[8] I. Redko, A. Habrard, and M. Sebban, "Theoretical analysis of domain adaptation with optimal transport," in *Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 737–753.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML'15*, vol. 37, 2015, pp. 1180–1189.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *NIPS'17*, 2017, pp. 5769–5779.

[11] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *NeurIPS'18*, 2018, pp. 8559–8570.

[12] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. AMD*, vol. 7, no. 3, pp. 162–175, 2015.

[13] W. L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, p. 026017, 2017.

[14] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[15] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR'17*, 2017, pp. 7167–7176.

[17] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV'19*, 2019, pp. 1406–1415.